A Controlled Benchmark for Memory Mechanisms in Efficient-Attention Transformers

Robert Sneiderman
Independent
robbysneiderman@gmail.com

Abstract

Efficient-attention transformers trade direct context access for compute, and a family of memory mechanisms is offered to recover what restricted attention drops. Reported gains are hard to trust: papers vary the mechanism, the backbone, the optimizer, the token budget, and the scale all at once, and they report a single aggregate perplexity that averages over positions where local attention already suffices. This benchmark holds the backbone, data, optimizer, and token budget fixed across four memory mechanisms on a 286M-parameter GPT, varying only the memory system, and scores them with a position-resolved diagnostic rather than one aggregate number. The diagnostic is a *positional context deficit map*: the per-position bits-per-byte (BPB) gap between local and matched global attention, against which a mechanism's per-position gain is measured. Across 42 training runs (three context lengths, up to four seeds each, paired by seed), only persistent memory tokens improve over the local baseline (-1.79 milli-BPB, paired $p \approx 0.012$, Cohen's d large); recurrent memory, test-time-training, and gated DeltaNet each hurt under a fixed budget. The deficit map shows persistent memory closes more than the full late-position deficit at the two longest contexts. Effects are small and single-scale; the contribution is the controlled protocol and the diagnostic, both architecture-agnostic and released as a plug-and-play harness.

1 Motivation

Sliding-window, sparse, and mixed local-global attention cut the quadratic cost of full attention by restricting each token's direct access to context. Memory mechanisms (persistent memory tokens, recurrent memory, test-time training, linear attention with state) are proposed to compensate by carrying information across positions or segments. Two evaluation habits make it hard to read the literature for fast ML. Mechanisms are introduced against different backbones, optimizers, token budgets, and scales, so a reported win confounds the mechanism with its training recipe. And evaluation collapses to one aggregate perplexity or BPB, which averages over positions where local attention is already enough and positions where it is not, hiding whether a mechanism solves the access problem or just lowers overhead everywhere.

This benchmark fixes both. It freezes the backbone, data, optimizer, and compute budget across mechanisms so the only variable is the memory system, and it scores mechanisms by where in the sequence they help, using a position-resolved diagnostic tied to a matched global-attention reference. The target is the fast-ML question: under a fixed efficiency budget, which memory mechanism is worth its overhead, and where.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: All mechanisms at T=2048, paired by seed against the SSSL baseline. Negative Δ is an improvement (lower BPB). p-values are paired t-tests.

Mechanism	Mean BPB	Δ vs baseline	p
Baseline (SSSL, local)	0.84599	—	—
Persistent Memory	0.84420	-1.79 mBPB	0.012
Gated DeltaNet	0.84734	$+1.35$ mBPB	0.008
TTT-Linear	0.85198	$+6.00$ mBPB	< 0.001
RMT	0.88524	$+39.25$ mBPB	< 0.001

2 Benchmark design

Fixed substrate. Every mechanism is a `MemoryModule` plugged into the same 12-layer, 768-dimension GPT (nanochat backbone: GQA, QK-norm, RoPE with θ=10^5) at 286M parameters. The efficient-attention pattern is SSSL: three of every four layers use a short window ($\approx T/4$) and one uses full context. Each run trains for 2,520 steps (≈1.32B tokens) on FineWeb-Edu with Muon on the body and AdamW on embeddings, head, and memory parameters. Memory mechanisms add at most 0.2% parameters. A run takes 13–25 minutes on an 8×H100 pod, so the full benchmark is 42 runs in roughly 53 GPU-hours: cheap enough to put error bars on every cell.

Three reference conditions. *Local* is the SSSL efficient baseline. *Global* is the same model with full attention on all layers, an operational reference for unrestricted access, not an information-theoretic ceiling. *Memory* is SSSL plus one mechanism. Persistent memory replaces the attention module rather than augmenting SSSL, so its comparison isolates learned memory tokens against full attention.

Position-resolved scoring. BPB is bucketed into 64-token position bins. The *positional context deficit* is $\mathcal{D}_T(p) = \mathrm{BPB}_{\text{local}}(p,T) - \mathrm{BPB}_{\text{global}}(p,T)$, the per-position cost of restricting access. The *memory gain* is $\mathcal{G}_T(p) = \mathrm{BPB}_{\text{local}}(p,T) - \mathrm{BPB}_{\text{memory}}(p,T)$, and *closure* $\mathcal{C}_T(p) = \mathcal{G}_T(p)/\max(\mathcal{D}_T(p),\epsilon)$ is the fraction of the deficit a mechanism recovers (masked where the local model is already sufficient). Closure near 1 means the deficit is repaired; above 1 means the mechanism exceeds the global reference at that position. This separates two questions a single BPB conflates: does memory help somewhere, and does it help where the architecture is most impaired.

3 Results

One mechanism survives a fixed budget. At T=2048 over four seeds, only persistent memory beats the local baseline (Table 1); recurrent memory (RMT), test-time-training (TTT-Linear), and gated DeltaNet each raise BPB. The persistent-memory gain is -1.79 milli-BPB, significant under a paired t-test ($p \approx 0.012$) with a large per-seed effect size, but small in absolute terms (about 0.2% of baseline BPB) and from only three to four seeds per cell. The four mechanism-versus-baseline tests are not corrected for multiplicity, so the p-values are read as suggestive pending a larger replication. The negative results are informative, not failures of the harness alone: RMT's segment recurrence is architecturally mismatched to SSSL, and TTT and DeltaNet carry replacement parameters that cannot use Muon under the distributed setup, a conservative bias the benchmark reports rather than hides.

The deficit grows, and memory closes it where it is largest. The peak per-position deficit grows with context length (0.53, 0.59, 1.38 milli-BPB at T=$2048, 4096, 8192$): these are the largest 64-token-bin values of $\mathcal{D}_T(p)$, distinct from the sequence-averaged BPB in Table 2, where the local-minus-global gap is smaller (0.45, 0.58, 0.62 milli-BPB) because most positions carry no deficit. Persistent memory holds its ≈ -1.8 milli-BPB gain across all three lengths and beats both the local and the global reference at every length (Table 2). In the deficit map (Figure 1), the gain concentrates in late positions, and late-region closure exceeds 1.0 at T=4096 and T=8192: persistent memory recovers more than the full restricted-access deficit there. One reading, following work on attention sinks and register tokens, is that a few learned memory slots act as a compressed bottleneck; the benchmark treats this as a hypothesis the diagnostic surfaces, not a settled claim.

Table 2: Persistent memory versus local and global references across context length (3 seeds per cell at 4096 and 8192). Negative is improvement.

Context	Local (SSSL)	Global	Persistent Memory
2048	0.84599	0.84554	0.84420
4096	0.84712	0.84654	0.84529
8192	0.84932	0.84870	0.84763

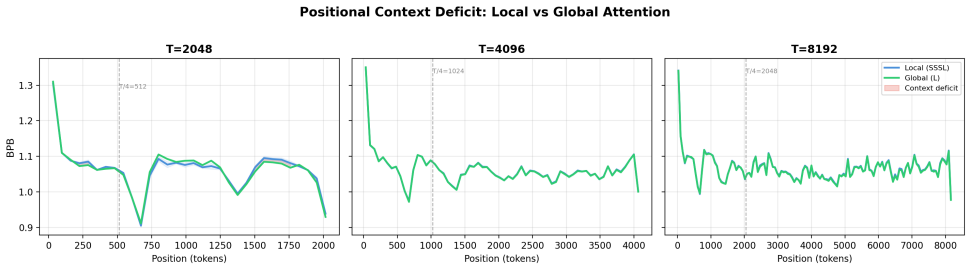

Figure 1: Positional context deficit map. Per-position BPB gap between local and matched global attention across context length; the deficit concentrates in late positions and grows with T. Persistent memory's gain (not shown) overlaps this high-deficit region, which is what the closure metric quantifies.

4 Scope and limitations

Three boundaries are explicit. Effects are small: the surviving gain is ≈0.2% of baseline BPB, statistically significant but of unknown practical value, and no downstream task evaluation has been run to test whether it transfers. The fixed token budget structurally favors the simplest mechanism, since persistent memory adds 32 static slots per layer while TTT and DeltaNet must learn recurrent dynamics; a fair extension trains those for 5–10× longer. And the benchmark is single-scale and single-metric (286M parameters, BPB only), so results need not transfer to larger models. The contribution is the controlled protocol and the position-resolved diagnostic, both released as a plug-and-play harness where a new mechanism is one file. The immediate next step is a downstream-task and longer-budget replication that tests whether the late-position closure observed here survives at scale.

AI usage

Prose was drafted and revised with a large language model, and the harness was co-developed with the same assistant under human direction. Reported numbers are read from the released run logs and statistical-test outputs, not generated by a model. Every citation was verified by the author against arXiv records. The author is responsible for all claims and code.

References

[1] Sukhbaatar, S., Grave, E., Lample, G., Jegou, H., Joulin, A. Augmenting self-attention with persistent memory. arXiv:1907.01470 (2019). `https://arxiv.org/abs/1907.01470`

[2] Bulatov, A., Kuratov, Y., Burtsev, M. Recurrent Memory Transformer. arXiv:2207.06881 (2022). `https://arxiv.org/abs/2207.06881`

[3] Sun, Y., Li, X., Dalal, K., et al. Learning to (learn at test time): RNNs with expressive hidden states. arXiv:2407.04620 (2024). `https://arxiv.org/abs/2407.04620`

[4] Yang, S., Wang, B., Zhang, Y., Shen, Y., Kim, Y. Gated Delta Networks: improving Mamba2 with delta rule. arXiv:2412.06464 (2024). `https://arxiv.org/abs/2412.06464`

[5] Liu, N. F., Lin, K., Hewitt, J., et al. Lost in the middle: how language models use long contexts. arXiv:2307.03172 (2023). `https://arxiv.org/abs/2307.03172`

[6] Darcet, T., Oquab, M., Mairal, J., Bojanowski, P. Vision transformers need registers. arXiv:2309.16588 (2024). https://arxiv.org/abs/2309.16588

[7] Jiang, A. Q., Sablayrolles, A., Mensch, A., et al. Mistral 7B. arXiv:2310.06825 (2023). https://arxiv.org/abs/2310.06825